CHALMERS
UNIVERSITY OF TECHNOLOGY

A Resource Sharing Method for Reliable Slice as a Service Provisioning in 5G Metro Networks

A Resource Sharing Method for Reliable Slice as a Service Provisioning in 5G Metro Networks

Elisabetta Amato*, Federico Tonini†, Carla Raffaelli*, Paolo Monti†
*DEI, University of Bologna, 40136 Bologna, Italy
†Department of Electrical Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden
E-mail: {e.amato,carla.raffaelli}@unibo.it and {tonini,mpaolo}@chalmers.se

***Abstract*—This paper proposes a dynamic slice provisioning analysis in a 5G metro network with reliability guarantees and possible sharing of backup resources. Performance of dedicated (DP) and shared (SP) protection solutions are evaluated with respect to slice resource allocation (i.e., bandwidth and processing units). The main results show a remarkable saving, in terms of slice acceptance rate, by applying SP solutions with respect to conventional DP ones.**

***Index Terms*—5G, slice as a service, reliability, shared protection, dedicated protection, latency, simulator**

I. Introduction

Network slicing allows the provisioning of different services over the same 5G infrastructure, where virtual or physical resources are interconnected to form end-to-end logical networks (i.e., the slices) [1]. Slicing allows service providers to offer 'network slices-as-a-service', tailored to different performance requirements [2]. When a slice is admitted by a provider (i.e., it is deployed over its infrastructure), it needs to be assigned a proper set of resources (i.e., connectivity and compute) to meet the Service Level Agreement (SLA) stipulated with the client [3].

In this respect, provisioning slices with very stringent reliability and latency requirements is an important challenge to tackle [4]. In the presence of failures, to avoid severe service interruptions while keeping the number of backup resources low, optimized provisioning of extra resources (dedicated or shared), is required. Compared to using dedicated backup resources, an approach leveraging on shared protection resources potentially leads to (i) fewer resources consumed by each slice, and, consequently, (ii) to a better slice admission ratio performance (or, equivalently, a lower blocking probability). These advantages come at the cost of a slightly longer recovery time (i.e., compared to using dedicated backup resources) due to the need to switch from the primary to the backup resources. In [5], [6], the authors evaluate the impact of different techniques for dedicated and shared backup protection on optical network resources. In [7], [8] efficient shared and dedicated protection schemes for cloud and baseband resources in 5G access/metro networks are applied. All the works mentioned so far consider only one technology domain (either transport or cloud), while [4] presents static resource provisioning strategies for dedicated and shared backup resources, considering transport and cloud domains jointly, for a single URLLC service slice provisioning.

This paper considers, on the other hand, the slice as a service paradigm. It proposes a heuristic that tackles the problem of dynamically provisioning slices with stringent latency and reliability requirements while minimizing the amount of transport and cloud resources assigned to each slice. The intuition behind the proposed approach is to encourage sharing of backup connectivity and cloud resources as much as possible. The performance of this shared protection scheme is compared against a conventional dedicated protection mechanism in terms of slice blocking probability and required processing resources considering a sample 5G metro network, where processing power is often limited and requires efficient resource allocation.

This work was supported by EUREKA cluster CELTIC-NEXT project AI-NET-ANIARA funded by VINNOVA.

II. Reliable Slice as a Service Provisioning

A. Architecture and Problem Formulation

In 5G networks, baseband functionalities can be virtualized over general-purpose hardware and centralized at different computing locations in the network, reaching different degrees of savings and performance targets [1]. The transport network links interconnecting the different compute nodes must be dimensioned accordingly to meet bandwidth and latency requirements for baseband and service processing. This paper considers a metro network comprising a set of source and target nodes connected by high-capacity optical transport links. The source nodes collect a set of antennas covering a given area and injecting traffic into the transport network. The target nodes are equipped with Processing Units (PU) to perform virtual baseband functions and services. The formulation of the dynamic and resilient slice allocation problem can be summarized as follows. **Given**: a network topology with available network resources (bandwidth and PU) and the requirements of a slice to be provisioned; **Find**: a suitable slice deployment, such that the allocated network resources are minimized; **To ensure**: reliability against single link or node failure (including the target node) while ensuring that the bandwidth and PU resources allocated at each link and node do not exceed the available resources, and that the maximum distance between a source node and a target node is enforced.

B. Methodology

To study the problem, an event-driven simulator written in Python was developed. Events consist of slice requests

978-3-903176-33-1

Algorithm 1 Slice Admission

1: **Algorithm:**
2: **for all** $primary_i$ **in** $Pair$
3: **if** $primary_i$ **meets** $slice\_requirements$
4: **for all** $backup_j$ **associated with** $primary_i$
5: **if** $backup_j$ **meets** $slice\_requirements$
6: **compute** $cost_{i,j}$
7: **add** $primary_i$ **and** $backup_j$ in $List_{pair}$
8: **if** $List_{pair}$ *is Empty*
9: $slice\_rejected$
10: **else ascending sort** $List_{pair}$ **based on** $cost_{i,j}$
11: **hold** $resources$ **required by** $List_{pair}[0]$
12: $slice\_accepted$

originating at random source nodes, which can be allocated or rejected in relation to the amount of resources available in the network. Each slice also has a lifetime after which the allocated resources are released. A pre-processing phase generates a set (referred to as $Pair$) of primary-backup path pairs between each source and target node. Each primary-backup pair is path disjoint and terminates at different target nodes. Each path is obtained using the k-shortest path algorithm with $k = 5$. Two different approaches for protection are considered, Dedicated Protection (DP) and Shared Protection (SP). Slice resource assignment is carried out as shown in Algorithm 1. For each possible primary path in $Pair$ (line 2)), the algorithm checks if latency constraint is met, and if there are enough resources (bandwidth and PU) available on both the primary path ($primary_i$) and at the candidate target node (i.e., to which the candidate primary path connects the source node to) to allocate the slice (line 3). The selection of the backup path and target node $backup_j$ (line 5) depends on the specific protection strategy. DP uses dedicated backup resources. As a result, a procedure identical to the one used for the selection of the primary path and target node is used. With SP, backup resources (bandwidth and/or PUs) can be shared at no additional cost if their respective primary paths and target nodes are disjoint. Otherwise, new backup resources are assigned. If enough resources are available on the considered primary/backup pair, and latency constraint is met, a cost ($cost_{i,j}$) is calculated as $\alpha * \sum_{i \in Links} Band_i + \beta * \sum_{j \in Nodes} Compute_j$, where α and β are coefficients used to balance the two contributions, $Band_i$ is the connectivity requirement (in Gbps) of the slice and $Compute_j$ is the PU required by the slice for the virtual baseband and the service (line 6). The value of $cost_{i,j}$ is saved together with the primary-backup pair (line 7) and, after exploring all possible pairs, the smallest cost pair is chosen (line 10). The resources needed for the slice are then booked into the network (line 11), and the slice is allocated. Resources are released after the expiration time. If there are no candidate pairs due to lack of resources, the slice is rejected (lines 8-9).

III. Evaluation

The network considered in the analysis is depicted in fig. 1. It consists of 14 source (blue) and 7 target nodes (red). The target nodes are chosen among those with nodal degree equal to 4, to allow better accessibility from the source nodes. The links in the transport are bidirectional, are assumed to

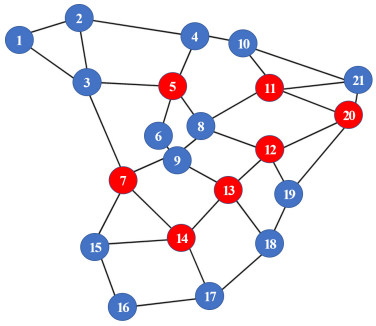

Fig. 1. Reference network with source (blue) and target (red) nodes.

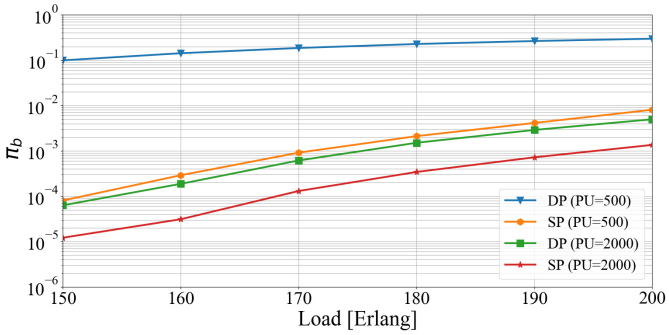

Fig. 2. Blocking probability as a function of the load.

have the same length, and have a capacity of 1000 Gbps. This study considers the deployment of slices with low latency and strict reliability requirements. The maximum number of hops allowed for a slice (to satisfy the latency requirements) is set to 4. The slice connectivity requirement is 24 Gbps and the compute one is 12 PUs (i.e., for baseband and service processing adopting split option 8 [4]). The distribution of the average inter-arrival frequency is exponentially distributed with $\lambda = 1$ per time unit. The average lifetime of the slice (θ) is also exponentially distributed, varied to consider different values of the network load ($A_0 = \lambda * \theta$). α and β are set to 1.

The results compare the two different protection schemes (DP and SP) in two network configurations, one with 500 PU (case 1) and the other with 2000 PU (case 2) available at each target node. The following quantities are introduced for evaluation:

$$B_S = \frac{\sum_{i=1}^{N} \frac{B_i}{S_i}}{N} \quad (1.1) \qquad PU_S = \frac{\sum_{i=1}^{N} \frac{PU_i}{S_i}}{N} \quad (1.2)$$

$$B_U^{j,k} = \frac{\sum_{i=1}^{N} B_i^{j,k}}{N} \quad (2.1) \qquad PU_U^{j} = \frac{\sum_{i=1}^{N} PU_i^{j}}{N} \quad (2.2)$$

where B_S represents the average bandwidth occupied by a slice in the network, defined as the ratio between the total bandwidth used in the transport network B_i and the number of active slices S_i when slice i is accepted, averaged over the number of events "slice accepted" N. PU_S represents the average number of PUs per slice, defined in (1.2) in a similar way, where PU_i indicates all the PUs used when slice i is accepted. The average bandwidth per link $j-k$ ($B_U^{j,k}$) and the average number of PUs per node j (PU_U^j)

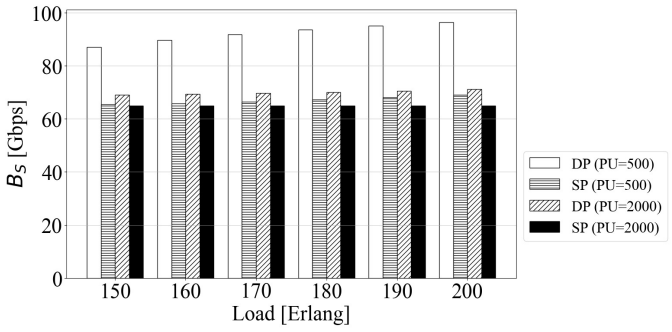

Fig. 3. Average bandwidth per slice (B_S) as a function of the load.

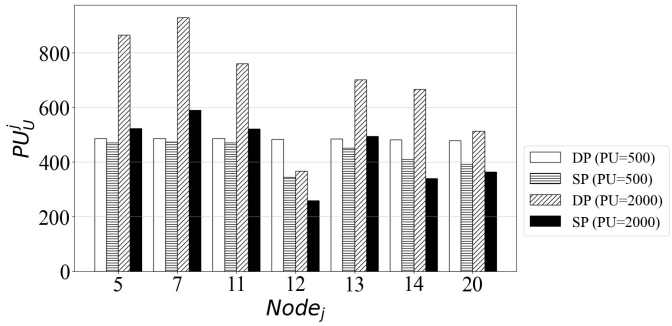

Fig. 4. Average PU per target node j (PU_U^j) with load = 200.

used in the network are represented by (2.1) and (2.2), respectively, where $B_i^{j,k}$ is the bandwidth allocated in the link that connects the node j and k, and PU_i^j indicates the PUs allocated in node j when slice i is accepted. The blocking probability is represented by (3), where N_r is the number of events "slice request", Bl_i is equal to 1 if the slice is rejected or 0 if it is accepted.

$$\pi_b = \frac{\sum_{i=1}^{N_r} Bl_i}{N_r} \tag{3}$$

Figure 2 compares SP and DP in terms of blocking probability. SP outperforms DP, in particular when PU resources are scarce (case 1). Figure 3 reports the value of B_S as a function of load. SP allows savings of up to 28% and 8.6% in case 1 and 2, respectively. While in case 2 there is almost no effect with different load conditions, in case 1 DP requires 9.6% additional bandwidth when passing from low to high load. This is because resources in the target nodes are scarce and saturates, forcing DP to try to reach nodes with available PU that are further away. This is shown in Fig. 4, where the average number of PUs per node is reported for load $= 200$. In case 1, DP uses on average all resources in all the nodes. SP is able to use resources more efficiently, thus the lower blocking probability. In case 2, the average PU usage is below 50% for both DP and SP. The reason for the higher blocking shown by DP is bandwidth availability over some links. This can be seen in Fig. 5, where the bandwidth used per link is, on average, close to saturation. Table I shows how many PU can be saved, on average and per slice, using SP compared to DP. As the load increases, the savings slightly increase, reaching 37% savings per slice. This is due to the higher number of

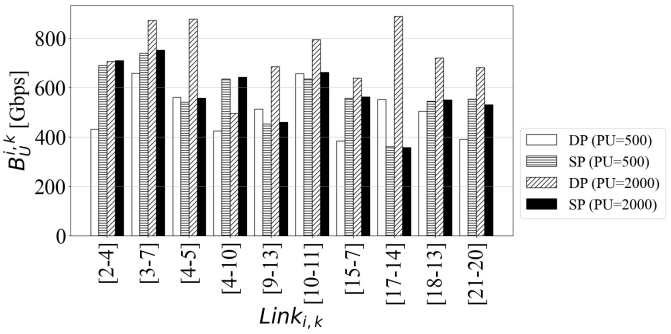

Fig. 5. Average bandwidth per link $j-k$ ($B_U^{j,k}$) with load = 200. Only links with utilization $> 60\%$ are reported.

TABLE I
AVERAGE NUMBER OF PUS PER SLICE: SP SAVINGS COMPARED TO DP FOR DIFFERENT LOAD VALUES.

Load	SP Savings (PU=500)	SP Savings (PU=2000)
150	35.62%	35.41%
160	35.88%	35.50%
170	36.16%	36.61%
180	36.47%	36.71%
190	36.75%	36.79%
200	37.03%	37.85%

slices activated simultaneously, which allows sharing a larger number of backup PUs.

IV. CONCLUSION

The paper presented a performance comparison between dedicated and shared protection schemes for dynamic slice provisioning in a 5G metro network context where processing resources per node are typically scarce. Results show that, especially in these conditions, the SP leads to considerably lower blocking probability than DP. SP is shown to save up to 37% PUs and 28% bandwidth per slice with respect to DP.

REFERENCES

[1] A. A. Barakabitze, A. Ahmad, R. Mijumbi, and A. Hines, "5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges," *Computer Networks*, vol. 167, p. 106984, 2020.

[2] X. Zhou, R. Li, T. Chen, and H. Zhang, "Network slicing as a service: enabling enterprises' own software-defined cellular networks," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 146–153, 2016.

[3] F. Tonini, C. Natalino, M. Furdek, C. Raffaelli, and P. Monti, "Network slicing automation: Challenges and benefits," in *2020 International Conference on Optical Network Design and Modeling (ONDM)*, 2020.

[4] F. Tonini, E. Amato, and C. Raffaelli, "Optimization of optical aggregation network for 5G URLLC service," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[5] N. Shahriar, S. Taeb, S. R. Chowdhury, M. Zulfiqar, M. Tornatore, R. Boutaba, J. Mitra, and M. Hemmati, "Reliable slicing of 5G transport networks with bandwidth squeezing and multi-path provisioning," *IEEE Transactions on Network and Service Management*, vol. 17, no. 3, 2020.

[6] A. Marotta, D. Cassioli, M. Tornatore, Y. Hirota, Y. Awaji, and B. Mukherjee, "Reliable slicing with isolation in optical metro-aggregation networks," in *2020 Optical Fiber Communications Conference and Exhibition (OFC)*, 2020, pp. 1–3.

[7] B. M. Khorsandi, F. Tonini, and C. Raffaelli, "Centralized vs. distributed algorithms for resilient 5G access networks," *Photonic Network Communications*, vol. 37, no. 3, pp. 376–387, Jun 2019.

[8] H. D. Chantre and N. L. Saldanha da Fonseca, "The location problem for the provisioning of protected slices in NFV-based MEC infrastructure," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 7, 2020.